BOSTON UNIVERSITY
SCHOOL OF EDUCATION
Ed.
LIBRARY
Thesis
Ed.D.
Read, J.G.
1946
The Gift of John G. Read
stored

To my wife, Helen G. Read,
who made it possible for
me to write this thesis
by her belief in my ability to
do so.

Ed.
Thesis
Ed.D.
Read, J. G.
1946
stored

BOSTON UNIVERSITY
SCHOOL OF EDUCATION

MEASURING, WITH A NON-VERBAL PICTURE-TEST, THE ABILITY TO RECOGNIZE AND CLASSIFY SOME OBJECTS AND SITUATIONS ASSOCIATED WITH THE PHYSICAL-SCIENCE ENVIRONMENT

Submitted by

John Gammons Read
(B.S., Massachusetts State College, 1924)
(A.M., Brown University, 1932)

In partial fulfillment of requirements for
the degree of Doctor of Education

1946

First Reader: William C. Kvaraceus, Ed. D.
Assistant Professor of Education

Second Reader: Donald D. Durrell, Ed. D.
Dean, Professor of Education

Third Reader: G. Lawrence Rarick, Ph.D.
Assistant Professor of Science Education

Boston University
School of Education
Library

School of Education
Gift of J. G. Read
Nov. 22, 1946.
27484

The actual Picture-test, a box 11 by 15 inches, is part of this dissertation.

Digitized by the Internet Archive
in 2016 with funding from
Boston Library Consortium Member Libraries

https://archive.org/details/measuringwithnon00read

TABLE OF CONTENTS

CHAPTER Page

I. The Problem 1

The test

Control of extraneous factors

Definition of terms

Organization of the study

II. Review of the Literature 5

Non-verbal testing

Tests of physical-science concepts

Aims of science teaching

Relation of this study to aims

III. Construction of the Test 14

Materials used

Criteria for choice of materials

Preliminary experiment

IV. The Scope of the Study 20

Design of the experiment

Collection of the Data

Administering the Picture-test

CHAPTER Page

V. Evaluation Techniques 28

Small-sample theory

Analysis of variance

Evaluation of the Picture-test

Distributions and Correlations

Data from three tests

Normality of Picture-test score distribution

Examining the hypotheses of the design

VI. Conclusions 57

Reliability of measurement

Validity indications

Uses for the Picture-test

VII. Suggestions for Further Study 63

Prediction possibilities

Teaching with the test

The Picture-test a prototype

Appendix A. Pictures 70

Appendix B. Preliminary experiment 73

Tables from which computations were made

Pupil score and data card

CHAPTER I
THE PROBLEM

The test.-- The problem is the building and evaluation of a test, completely non-verbal, attempting to measure the ability to recognize and classify some objects and situations associated with the physical-science environment; using the matching of photographs to obtain an individual's score. The test consists of a box with a place for cue-pictures, in groups of four; bins beneath each of the four pictures; and movable single photographs (item-pictures) each of which is to be dropped into the bin which the subject thinks is the best placing for it. Science areas selected are: [1]

Practice: farm, factory, construction, retail store

1. uses of electricity
2. tools and applications of four formal sciences
3. molecular physics
4. electronics
5. chemistry
6. mechanics

---

[1] Pictures, Appendix A, page 71 .

Control of extraneous factors.-- Some of the factors which may tend to contaminate the results of tests are partially or wholly eliminated by the choice of design of the test. As there are no words used, all language and reading difficulties are nullified; directions for taking the test may be given in pantomime. Within each of the six parts of the test, the pupil has complete freedom in the making of changes in the placement of items until he is satisfied that they represent his best judgment; there is not even the psychological barrier of erasing a mark made to indicate a choice. After all items are placed, each of the four bins may be searched to see if all items have common attributes; changes again may be made without penalty. Except where motor difficulties are peculiar to some individual, there are no motor skills involved and there is no time limit to accentuate small variations in such skills. The vocabulary-difficulty has been eliminated. Pupils in a verbal test must respond to words, often failing to recognize them as representing realia which they really do know the use of, and could place in a larger category or concept. Also, students sometimes are able to place words in their proper categories, knowing neither the meaning of the word nor of the category. This spurious placement is found in dealing with concepts and generalizations in physical science.

The concepts chosen for this study are examples. They are stated at the beginning of Chapter III.

Definitions of terms.-- Of the terms used above, concept and generalization need definition for this study. As used in science education they are nearly synonymous. Hereafter, concept will be used to represent the idea of a comprehensive area of science composed of elements which have common scientific characteristics. The word principles is used by some of the authors quoted in much the same way. Examples of concepts in science given in the 31st Yearbook [1] are:

> The sun is the chief source of energy for the earth.
> All life comes from life and produces its own kind of living organism.
> Distances in space seem extremely vast when compared with distances on earth.
> All matter is probably electrical in structure.
> Liquid or gas pressure is exerted equally in all directions.

Organization of the remainder of the study.-- The chapters to follow will:

a. List the concepts and their subdivisions which were chosen for the test.

b. Review the literature for similar studies and for statistical design.

c. Describe the testing procedure and the groups tested.

---

[1] "A Program for Teaching Science", Part I, The Thirty-First Yearbook of the National Society for the Study of Education, Public School Publishing Company, Bloomington, Illinois, 1932, p. 53,54.

d. Give a summary of results of the testing and draw conclusions from them.

e. Give suggestions for further study using this test or similar devices.

CHAPTER II

REVIEW OF THE LITERATURE

Non-verbal testing.-- Among the tests reviewed were some having close relationships with the test of this study.

The Healy test,[1] used as part of the battery which comprised the Army Beta in 1917-18, has small movable square cards on which there are line drawings of objects. These can be made to fit into square holes in a larger line drawing showing a situation from which a key object is missing, which object should fill the hole. An example: boys holding a broken kite-string wound on a stick are looking up into the sky, where there is a square hole in the picture. The small movable cards offer drawings of birds, a balloon, a kite, an airplane, a baseball, and a plain white area. Similar cause-and-effect relationships were used in eleven items. This test has currently been used by Goldstein and Scheerer [2] at Tufts Medical School in testing brain-injured and defective children. These investigators have also used a collection of objects which

---

[1] W.Healy, "A Pictorial Completion Test", Psychological Review, Volume 21, 1914, p.189-203.

[2] Kurt Goldstein and Martin Scheerer, "Abstract and Concrete Behavior", Psychological Monographs, Volume 53,1941,p.1-149.

Halstead[1] arranges as a test. They are capable of being sorted into categories not necessarily unique. A chocolate cigar, a pipe, tobacco, and matches belong in one group; the cigar and the tobacco also may be grouped because both are brown. The purpose, says Goldstein[2] is to "survey objects actively with the purpose of generalizing". Close observation and questioning by the tester as to the purposes of the subject in making the classification are valuable, and produce more valid results than mere statistical scores. The Dodd International Group Mental Test[4] is described by Edds[3]:

> The International Group Mental Test by E.A.Doll [sic] is an arrangement of pictures into many grouped series such that an item in one series matches in some specific way an item in the other part of the same series. The more advanced sections of the scale have a centrally located picture which serves as a cue to similarity required. The subject matches the items by joining them with a pencil line.

One section of a test by Brody[5] is composed of

---

[1] W.C.Halstead, "Preliminary Analysis of Grouping Behavior in Patients with Cerebral Injuries", American Journal of Psychiatry, Volume 96, 1940, p. 1-64.

[2] W.C.Halstead, ibid., p.17.

[3] Jess H. Edds, "The Nature of Verbal and Non-verbal Abilities", Journal of Experimental Education, Volume III, No.3, March, 1935, p.225.

[4] Stuart Dodd, The International Group Mental Test, Doctorate Thesis, Princeton University, 1926.

[5] Leon Brody, "Comparable Tests of Verbal and Non-verbal Reasoning", Journal of Educational Psychology, Volume 31, No.4, April, 1940, p. 304.

pictures of common objects which are to be matched. He reported that this test had valid and reliable coverage from grades four through twelve. The whole test, attempting to measure four types of reasoning, (non-verbal concrete, non-verbal abstract, verbal concrete, verbal abstract), is summarized as follows:

> As measured by these tests, the four types of reasoning do not individually seem to involve the same things measured by ordinary, verbal intelligence tests. Verbal concrete and verbal abstract combined approach the latter to a significant extent, probably because of common verbality. Non-verbal concrete and non-verbal abstract do not. This does not mean that the verbal tests are better measures of intelligence as such, than the non-verbal tests. It may be that the non-verbal tests are getting at another aspect of intelligence.

In the manual of one of the Durrell-Sullivan tests [1] the section on "Paragraph Meaning" involves concept recognition:

> This test consisted originally of twenty paragraphs graded in difficulty from simple narrative written on a primary level to extremely difficult paragraphs as judged by vocabulary, sentence structure, and concept difficulties....Both factual and informational topics as well as selections from the field of science and history were used....the topics were designed to catch and hold the interest and attention of youngsters, so that the test would measure comprehension of the material rather than the degree of attentiveness the child gave to the reading of the paragraphs. Five questions were constructed to measure the comprehension of each of these paragraphs....Multiple-choice picture answers were provided for each of these questions. Great care was used in drawing the non-stimulus pictures so that careful discrimination would be required on the part of the child to pick out the correct answer.

---

[1] Donald D. Durrell and Helen Blair Sullivan, "Manual for Primary and Intermediate tests, Durrell-Sullivan Capacity and Achievement Tests",World Book Company,Yonkers-on-Hudson, New York, 1937, p. 3,4.

A test adapted for use by the armed forces in 1945 is composed of items from a test by Bennett [1]. It contains one section on mechanical comprehension, testing physics concepts to be deduced from diagrams.

Tests of physical-science concepts.-- While some of the tests mentioned above have science concepts included in their sub-tests, they do not seem to be attempting to measure the same ability or abilities which the writer's test is trying to measure. Some studies were discovered which do attempt to measure recall, reasoning ability, recognition, classification, and in general the getting of meaning from objects and situations of the physical-science environment. Smith[2] reports the results of using a test comprised of items designed to measure recall and reasoning ability with reference to measured intelligence. He finds that:

> 1. The abilities measured by the mental test and indicated by the I.Q. are not necessarily the major abilities involved in learning the general science material tested.
> 2. Ability to recall information and ability to see relationships between information and related principles are closely correlated.
> 3. While it is not possible to say that ability to recall information is sufficient to insure understanding of relationship, or vice versa, the two seem to be products of the same learning process and are highly and closely related to each other.

---

[1] George K. Bennett, Test of Mechanical Comprehension, Form AA",The Psychological Corporation, 522 Fifth Ave., N.Y., 1940.

[2] Victor C. Smith,"A Study of the Degree of Relationship Existing between Ability to Recall and Two Measures of Ability to Reason", Science Education,Volume 30, No.2, March 1946. p.88,89.

> 4. It is unlikely that a type of learning which fails to produce the ability to recall facts will be able to produce the ability to see relationships and to understand scientific principles.

Powers[1] found that for prediction of success in chemistry that:

> Terman test scores are as good as scores on the chemistry [achievement] test....This is an anomalous finding and entirely unexplainable by the investigator, unless it be said that both values of *r* are so low that chance factors have determined them.

The ability to apply learned facts to new situations is one of the stated aims of many science curriculum studies. Beauchamp and Webb[2] state as conclusions after a study of the resourcefulness of pupils in setting up apparatus, with non-standard materials, in the chemistry laboratory:

> If resourcefulness is really a desirable factor of a pupil's ability, then we are not justified in measuring him *only* by tests that do not measure this quality. If experience is one of the foundations of resourcefulness, then good training in the schools should include many experiences with actual objects, in varied fields of activity, related as frequently as possible to actual situations of practical living. The solution of simple perplexities gives training for more complex dilemmas.

Reasoning ability, the ability to group facts into categories, areas, or concepts having meaning; resourcefulness in using realia or approximations of realia, (as

---

[1] S.R.Powers, "The Correlation between Measures of Mental Ability and Measures of Achievement in Chemistry", *School Science and Mathematics*, Volume 28, December, 1928, p.981-986.

[2] Robert O. Beauchamp and Hanor A. Webb, "Resourcefulness, an Unmeasured Ability", *School Science and Mathematics*, Vol. XXVII, May, 1927, p.465.

are photographs); all seem to the writer to be sufficiently closely related to be considered *educationally* allied. To call them, as a group, problem-solving acticities, and to consider problem-solving the major activity of science education seems defensible. If further definition of the activity is necessary, it may be found in this quotation from Bode:[1] "This process of discovering and testing meanings....is what constitutes thinking."

Aims of science teaching.-- Some of the aims of science teaching that stress the acquisition of the scientific method of thinking should be criteria for any science test. Six aims synthesized from many sources by Noll [2] were:

> 1. Accuracy in all operations including calculation, observation and report.
> 2. Intellectual honesty.
> 3. Open-mindedness.
> 4. The habit of looking for natural causes [or the] conviction of universal operation of the law of cause and effect.
> 5. The habit of criticism,....particularly the ability and willingness to criticize oneself.

In part, these are the scientific attitudes based upon an analysis of the philosophical thought of Karl Pearson, John Dewey, and others, by Curtis:[3]

---

[1] Boyd Henry Bode, *Conflicting Psychologies of Learning*, D.C.Heath and Company, Boston, 1929, p.280.

[2] Victor H. Noll, *Teaching of Science in Elementary and Secondary Schools*, Longmans, Green and Company, Inc., New York, 1939.

[3] F.D.Curtis, *A Synthesis and Evaluation of Subject Matter Topics in General Science*, Ginn and Company, Boston, 1929, p.48.

1. Conviction of universal basic cause and effect relations....
2. Sensitive curiosity concerning reasons for happenings....
3. Habit of delayed response....
4. Habit of weighing evidence....
5. Respect for another's point of view....and willingness to be convinced by evidence.

A third list is to be found in Keeslar's article [1] as he reports the evaluation by 22 research scientists at the University of Michigan of 54 major and minor elements of the scientific method. After the subsequent abbreviation of the list they were:

1. Sensing a problem and deciding to try to find the answer to it.
2. Defining the problem.
3. Studying the situation for all facts and clues bearing upon the problem.
4. Making the best tentative explanation or hypothesis as to the best possible solution of the problem.
5. Selecting the most likely hypothesis.
6. Inventing and carefully planning one or more experiments to test the hypothesis, isolating the experimental factor wherever possible by using a control.
7. Testing the hypothesis by carrying out the experiment with great care and accuracy.
8. Running check experiments involving the same experimental factor to verify the results secured in the original experiment.
9. Drawing a conclusion.
10. Making inferences based on this conclusion when making new situations in which the same factors are operating.

Relation of this study to aims of science education.--

It is evident that as we attempt to measure the "ability to

---

[1] Oreon Keeslar, "The Elements of Scientific Method", Science Education, Vol. 29, No.5, December 1945, p. 275.

recognize and classify objects and situations associated with the physical science environment", we shall be measuring simultaneously the acquisition and retention of some facts which have been learned in school, and the ability to apply these facts to new situations. These situations are often vicarious, as in this test where they are pictured by means of photographs. Other facts have been learned from the out-of-school experiences of the pupil. There is extreme variation among pupils here, and only by providing several common categories will there be even a reasonable chance of measuring this awareness of an environment; there is little chance of testing every environmental situation in a measure of any reasonable length. This extreme range in experiential background must be present in all but those tests which have, as their basic forms, nonsense symbols, abstract designs, and similar devices such as Cattell used and called "culture-free". 1/ These do not measure specific concepts in the sense of this study, and are usually intelligence tests. The use of these abstract forms poses unique problems. From another angle it may be safely said that all tests of both picture and verbal types will be contaminated

---

1/ Raymond B. Cattell, "A Culture-free Intelligence Test,I", Journal of Educational Psychology, Vol. XXXI, No.3, March 1940, p.161.

with the socio-educational background variability mentioned above. It is the writer's opinion that no amount of equating of pupils will remove this factor. The only recourse is to use many extremely long and extremely diversified measures of aspects of the environment, and present objects and situations from those aspects so that they are in familiar guise. It would seem that finally a pattern of the subject's residual awareness could be discovered. This residual learned ability to recognize objects and situations would of course have been limited both quantitatively and qualitatively by the aptitude of the pupil. Inasmuch as the schools state that one aim of (science) education is to provide opportunities for the acquisition of increments in the ability to see the meaning of things in their natural settings, it is important that a direct measure of this ability be discovered. Such a test should give the pupil an opportunity to exhibit his increments in the abilities mentioned above with the least interference possible from other factors, which may be skills and concepts which are considered desirable by the school, but which nevertheless may not be indispensable in the good citizen in a democracy, living a worthy and happy life. While all of these advantages are not claimed for the writer's test, it does not violate any of the principles of this paragraph.

CHAPTER III

CONSTRUCTION OF THE TEST

Materials used.-- The six concepts chosen for testing were listed on page 1. Recasting them in the form of the concepts of science education they appear as follows: Practice: Each of four economic areas, (farm, factory, construction, retail store) have objects associated with them which are peculiar to one of those areas.

1. Each of four formal sciences (astronomy, geology, physics, chemistry) have objects associated with them which are peculiar to that science.

2. The uses of electricity may be divided into four categories (power, communication, heat-and-light, radiation as in X-rays) and there are associated objects peculiar to each category.

3. Four types of molecular activities of fluids (compressed gases can do work , atmospheric pressure in conjunction with other lower pressure can do work, expanding gases absorb heat which is later radiated as they are condensed, rapidly-moving fluid streams reduce pressure at the side of the stream) each of which is represented by objects peculiar to that type.

4. Four electronic applications (transmission and transformation of alternating current, production of direct current by chemical action, creation of potential differences due to accumulation of static charges, continuous controllable electron-emission due to heat and potential difference) are represented by objects peculiar to each application.
5. Four areas of chemistry (uses of petroleum products, uses of carbon dioxide, fats are one ingredient of soap, fuel in gaseous form supplies heat) are represented by peculiar objects associated with each area.
6. Four areas in mechanics (cantilever construction,simple harmonic motion, inclined planes or screw-threads, one type of lever) have peculiar objects associated with each area.

The ability to recognize these concepts as distinctive, and the ability to discriminate among each of the four subdivisions, constitutes the test. Photographs of the test, of pupils taking it, and of correct placement of each item will be found in Appendix A, page 71 .

Criteria for Choice of Materials.-- The subject-matter and the form of the test were chosen after many experiments with other devices, most of which involved the use of writing instruments or of styli. The writer feels that for pupils in a school situation the cramped tension which comes with writing is a possible bar to free association of objects or pictures. There is no objective proof at this

time for such a stand. However, the form of the test was chosen to eliminate any such factor. The test further allows opportunities for suspended judgment, for the revision of choices without even the psychological penalty of erasure. It furnishes high "game" type motivation, eliminates completely the language and reading factors, and eliminates much of the opportunity for giving or obtaining help from others taking the test. There is also perhaps some advantage of the sort that is discussed by Korzybski 1/, although this must wait for confirmation by a later study. Korzybski has found that in semantic therapy the labels for things (in his work, words printed on cards) become the things themselves to the subjects if they can handle and move the cards, and attach them to another card which represents other objects or concepts. It was this work which suggested the using of individual movable photographs to the author. The pictures were taken with this identification-factor in mind. Each one must have a clear-cut reason for placement in one best position. Further, this reason must be the unique and complete science link of the item-picture with the cue picture. The test has been examined by eleven science teachers and able laymen, and all items for which any other reason could be found for correct placement were removed. The same

---

1/ Alfred Korzybski, Science and Sanity, The International non-Aristotelian Library, Publishing Co., The Science Press Printing Company, Scranton, Pa.,1941.

process was carried out with thirty-one children at grade levels eight to twelve, and all items removed for thich they could find a reason which was not the unique scientific reason desired by the writer. More than 30 items were thus eliminated. Those remaining can be defended on the basis of exact and complete correspondence with one and only one cue-picture in each concept division. This is not to say that the reasons for incorrect placement of some items are not as revealing as are the correct ones. Individual use of the test while the subject talks of his choices is extremely enlightening as to the thought-processes employed. In order not to introduce factors of misunderstanding of the meaning of the item, the photograph was taken with extreme care so that the central object would be distinct and recognizable and would appear in its natural environment if that environment is significant. The eye-level is that of the high-school pupil, and the psychological point of view is that of the casual pedestrian or automobile passenger, seeing these objects in passing. The objects therefore must be reasonably accessible to many individuals in those parts of the United States where small cities and the countryside are available to young people, and where magazines and newspapers are in common use. The concepts chosen are those of everyday life, and are roughly apportioned among the physical sciences in the

following ratios: physics, 62, chemistry, 23, geology, 8.5, and astronomy, 6.5 per cent. These ratios are not significantly different from those found by Searle and Ruch [1] in a ten-year analysis of science articles in periodicals, nor from those discovered in an analysis of principles at many grade levels by Wise.[2]

Preliminary experiment.-- The Sullivan [3] "special fields" test suggested a similar device for science students in a teachers' college. (Her test had predicted with an extremely high reliability the success of student teachers as measured by the average of several independently reliable rating scales by competent instructors.) Six hundred science words were chosen by a sampling technique, and arranged first in order of difficulty. A rough item analysis cut the list to 400, and the test was given over a period of three years to six classes, once at the beginning and once at the end of the semester's science course which was of the survey type and included all the physical sciences. All students had previously completed a required course in biology in their freshman year. Correlations with final grades in the

---

[1] Albert H. Searle and Giles M. Ruch, "A Study of Science Articles in Magazines", School Science and Mathematics, Vol. 26, April, 1926, p.389-396.

[2] Harold E. Wise, "A Determination of the Relative Importance of Principles of Physical Science for General Education", Science Education, Vol.25, No.7, December, 1941, p.371.

[3] Helen Blair Sullivan, "A New Means of Appraising the Qualifications of Prospective Teachers", Harvard University, Graduate School of Education, unpublished thesis, 1944, 185p.

physical science course were not significant, and the war interrupted the attempt to trace the testees into their science-teaching positions in the schools. The experiment still awaits completion (1946). A copy of the test is found in Appendix B, page 74. This was an attempt to make classifications after recognizing the words which might represent realia to the subject as he read them.

CHAPTER IV.

THE SCOPE OF THE STUDY

Design of the experiment.-- The sampling of a population should be such that the sample will represent as nearly as possible that population which later will be the field for general use of the test.[1] This study attempts to sample grades 8 through 12 in the schools of areas around Boston, Mass., and Providence, R.I. Grades 11 and 12 were sampled so as to produce two groups. The first consists of pupils who have had chemistry or physics, or both. The second is made up of pupils who have had neither. These two groups will be called hereinafter the maximum-science group and the minimum-science group respectively. Some possibilities of differentiation by the test can be stated as null hypotheses.

1. There is no difference between the mean score of boys and the mean score of girls.

2. There are no differences among mean scores of the grades 8 through 12.

3. There is no difference between mean scores of pupils in the same grade but coming from places over 25,000 population and those coming from places under 7,000 population.

---

[1] E.F.Lindquist, Statistical Analysis in Educational Research, Houghton Mifflin Company, Boston, 1941, Chapter I.

The first group is designated as urban, the second as suburban.

4. There is no difference between mean scores made by those who have had maximum science and those who have had minimum science.

Seven places and twelve schools were selected. They represent a fair cross-section of the schools in the area whose foci are Boston and Providence, and all schools are within the sphere of influence of one of the cities. In the larger places (over 25,000) the city's influence is most marked. Pupils often go into the metropolitan area for recreation and study. The smaller places (under 7,000) are more rural, and in these schools there is a large proportion of students from homes where farming is at least a part-time source of income.

Fisher and Yates [1] have compiled tables of random numbers. If columns of these numbers are used to draw from class-lists a group of the required size, that group will be a truly random sample of the whole class. Care must be taken not to get an atypical sample in schools where ability-grouping is practiced. Randomized proportional samples from all groups must be obtained to make up the composite.

---

[1] R.A.Fisher and F.Yates, "Statistical Tables for Biological, Agricultural and Medical Research", Oliver and Boyd, London,1938, reproduced in part as Table 18, Lindquist, op.cit., p.262-264.

In this experiment each sample consists of 12 or 24 pupils, equally divided as to sex. (There are a few odd-numbered groups, later reduced to standard size by random withdrawal of a few cases.)

Pupils in grades 8 through 10, plus the minimum-science group of grades 11 and 12, were tested in an attempt to show growth in the trait being measured by the writer's Picture-test. The maximum- and the minimum-science groups in grades 11 and 12 were expected to furnish considerable information about the Picture-test, inasmuch as it is concerned with science facts and concepts. High scores should be found in those groups containing pupils who have studied in school these higher levels of physical science labeled chemistry and physics. While all members of the maximum-science group had not taken both subjects, it was felt that interest, ability, and future needs had been factors deciding membership in an advanced science group. This group, then, was highly selected as to the factors just mentioned.

In addition to the randomly-selected pupils a special group,called the Talent Group in the tables,was selected. It consisted of eighteen pupils from grades 8 through 12 with thirteen boys and five girls included. They all were members of a Science Club in a high school and were taught by an able and unusual teacher. Membership in the club is

elective, with the approval of the instructor, to those who have had good marks in science for at least the preceding semester. The Club meets daily and its 25 members produce realia which solve problems in science. These projects are used in classes as demonstrations of science principles. The abilities demonstrated by the members of the club are allied to those which the Picture-test is attempting to discover.

This Talent Group entered the state-wide Westinghouse Science Talent Search, one phase of which was a competitive Science Fair for the exhibition of required projects. There were 250 entrants, and 70 First Science Grants were made. Of these first awards, the Science Club mentioned above took 18, leaving 52 to be divided among 21 other schools.

Collection of the Data.-- Because the background knowledge of physical science was one determining factor of the scores on the Picture-test, it was necessary to discover how much this affected the results obtained from the administration of the Picture-test. The Cooperative General Science Test, Series P,[1/] for grades 8 through 12 was given to all pupils. This was a particularly appropriate test, as there are separate norms for those who have had chemistry

---

[1/] O.E.Underhill and S.R.Powers, Cooperative General Science Test, American Council on Education, Cooperative Test Service, New York City, 1939.

and physics, and these are so closely parallel that it is evident that the writer's plan,(of considering pupil-membership in either advanced science sufficient to place him in a special category), is reasonable. Intelligence quotients from school record cards were used if the latest record was not more than three years old. It may be helpful to experimenters to mention that there were "spare" pupils tested in each group, so that absences, defective scores, or conflicting data of any kind could eliminate a pupil without spoiling the design of the study. The chronological age for each pupil was also obtained; with the I.Q., the general science achievement-test score, and the Picture-test score, there are four statistics for each pupil.

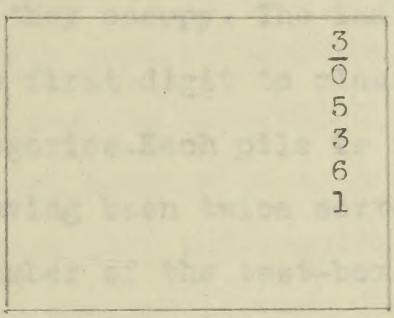
Administering the Picture-test.-- A preliminary visit to the school was made. Pupils who were to take the test later were randomly-selected, a quiet room with adequate light (measured with a foot-candle meter) chosen, and the science teacher visited. Achievement-test blanks were left with him with directions as to their administration within one week after the giving of the Picture-test. All Picture-tests were given by the writer in order to minimize the factor of teacher-pupil relationship.

On the day of the test, pupils entering the empty test-room found the closed test-boxes (see pictures of test and its administration in Appendix A, page 71)on chairs or desks.

The record-card was filled out (Appendix B, page 81) and card and pencils put aside. It is probably important that both hands be completely free during the taking of the test. The directions for giving the test are:

> Do not open the box until I tell you to. You are going to take a science test. There will be no written work, and there is no time limit on the test. Take all the time you wish.
>
> In the box you will find some pictures arranged on a blue card like this. (Show sample.)
>
> There will also be some separate, loose pictures like these. (Show samples.)
>
> Each one of the separate, loose pictures can be matched best with one of the pictures on the card. Here are some practice pictures. (Show that they match with the four pictures on the card; then show the dairy farm.)
>
> This one might go in any place. (Let them choose, but show that it can fit in any one of the four places, but that the best place is with the cow.)
>
> Open the box. Pick up the long card that is near you. Place it (with the pictures face-up) in the space above the little boxes. (Demonstrate.)
>
> Now pick up one loose picture. Place it in the little box under that picture which you think it matches best. Do this with all the pictures. Take all the time you wish.
>
> (When they get near the end of the practice area.) Place the blue papers, one on top of each pile of pictures in the little boxes, when you finish the first part of the test. Then turn over the next card. Go right on with the test until you finish. Remember that there is no time limit on the test.

After the test is completed, the record-card is placed inside the box and the cover closed firmly. This automatically covers the small boxes or bin, now filled with pictures, and allows the test to be transported without disturbing the pupil's placement of pictures. The test may be scored in about four minutes, using the following method. On the back of each individual picture there is a code-number, like this:



The 3 indicates the sub-test to which the item belongs; 305 is the number of the item (see Appendix B ,page 77 for list of items and their numbers). There remain three digits, two of them dummies, always either 1, 2, 3, or 4; and one of them, the key digit, either 5, 6, 7, or 8, representing the four bins numbered from left to right. In scoring, all the item pictures are taken from the extreme left-hand bin, where they were placed by the pupil. They are turned over, and sorted. All those that have a 5 appearing as one of the last three digits are correctly placed, and are put aside. The remaining pictures of this group are errors, and they are recorded on the pupil's

record card under column I, which indicates that they came from the first bin. Errors are self-checking, for they are recorded and the number is again noted;then the correct items are searched again to see that no error has been allowed to escape. This process is followed for each of the four boxes or bins, the total errors are added and the score determined. The record card is again checked against a master-list of numbers to make sure that the errors recorded are errors in the position which they occupy. The individual items are now re-sorted,using the first digit to classify them into the seven sub-test categories.Each pile is returned to its proper place in the box,having been twice sorted and finally shuffled. On each card the number of the test-box is placed by the pupil. Each evening the box is checked for completeness to make sure the test has not lost a picture.

The final card appears as illustrated (Appendix B,page 81) and contains all the data gathered concerning the pupil. Because these cards may be combined into many different categories, the statistical examination of the test is facilitated. The technique just described approximates some of the Hollerith punched-card techniques and does away with the necessity of much tabulation.

CHAPTER V

EVALUATION TECHNIQUES

Small-sample theory.-- In general, the theory of estimation, from small samples, of population characteristics, rests on two assumptions. The first is that random sampling, used in conjunction with the proper modifications of formulas, will accurately predict the mean and the standard deviation of the population from which the small sample was drawn. The second assumption is that while we never can be absolutely sure that our results are exact, we can determine how often, say out of each 100 samples like the one we have drawn, the statistics which we have obtained and which we wish to apply to the whole population, could have arisen by chance alone. In small samples, the use of the normal probability table, based on the areas cut off under a normal probability (Gaussian) curve, is not proper, because small samples tend to be markedly peaked; therefore another statistic, known as Student's t, is used to determine the level of confidence which we can have that any particular result which we obtain arose only by chance.[1]

The sampling distribution of t is known to be independent of all except one factor, the number of "degrees

---

[1] E.F.Lindquist, Statistical Analysis in Educational Research, Houghton Mifflin Company, Boston, 1940, p.51.

of freedom"[1] which itself is a function of the number of cases and of statistics concerned with the number of cases. The t-ratio is used frequently to determine the probability that a difference between means was a chance result alone. Tables give the level of confidence associated with a particular value of t for some degree of freedom. The two levels most used are the 5 per cent and the 1 per cent levels of confidence.

If we use the formula for the significance of the difference between two means derived from related measures (i.e. having the same variance or standard deviation squared) which is [2]

$$t = \frac{M_1 - M_2}{\sqrt{\frac{\sum d^2}{n(n-1)}}}$$

and where M_1 and M_2 are the means, $\sum d^2$ is the sum of the squared deviations of all the scores from the mean of the scores, and n is the number of cases, and thus find this t to be 3.169 for 11 cases (10 degrees of freedom) we are confident that only once out of 100 times (1 per cent level of confidence)could our sample, by chance,have had so great a difference between means. We can also have a great deal

---

[1] R.A.Fisher, Statistical Methods for Research Workers , Oliver and Boyd, Edinburgh, 1925.

[2] E.F.Lindquist, op.cit., p.54.

of confidence, (i.e. at a known level of probability) that the difference in means is due to some causal factor, but what that factor is, and in what direction it is operating, our t statistic does not tell us. Lindquist [1] derives the formula for the t ratio when the difference between two unrelated means is involved, using the variances of two populations:

$$t = \frac{M_1 - M_2}{\sqrt{\left(\frac{\sum d_1^2 + \sum d_2^2}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

The same statistic t can be used to test any null hypothesis which may be postulated about small random samples. Values of t should not be confused with those of the critical ratio as the C.R. is, unlike t, concerned with the normal probability curve.

It is frequently desirable to know whether two means may be compared for significance. We need to ascertain whether they come from two different populations; or whether, in fact, they are the means of two groups which should really be considered so far as their variance is concerned, as two parts of the same population. The F-test is often used to determine whether there is a significant

---

[1] E. F. Lindquist, op. cit., p.57.

difference between two variances. This statistic is

$$F = \frac{\sigma_1^2}{\sigma_2^2}$$

or the variance ratio (with the larger variance in the numerator of the fraction) of the estimated true population variances, each derived from

$$\sigma^2 = \frac{\sum d^2}{n-1}$$

Lindquist [1] after Snedecor [2] has tables similar to those for t for evaluation of the F-ratios. It is the usual practice to test the variances of two populations for significance before applying the t-test. When t is used, it is assumed that the samples are random samples from a normal population; that is, the variances of the groups used should be about the same. However, even though the distribution is not quite normal, the t-test is reasonably valid.[3] It should be stressed again that populations are like or unlike each other depending on two characteristics, their means and their standard deviations. To repeat, the latter statistic often appears in the squared form as the variance.

---

[1] E.F.Lindquist, op.cit., p.60-68.

[2] G.W.Snedecor, "Statistical Methods", Collegiate Press, Inc., Ames, Iowa, 1937, Table 10.2, p. 174-177.

[3] E.F.Lindquist, op.cit., p.71.

There has now been described a method of comparing two or more groups, using their means and variances as indices of the group achievement on a particular criterion or test. There are two conclusions possible: first, that the groups are members of the same population, and any observed variability in means and variances is due only to chance fluctuations: second, that they are not members of the same population, our statistical search having ruled out all of that part of the variability which is due to chance, leaving a residue of variability of response from group to group which is greater(at some predetermined level) than that which may be ascribed to chance. Note again that there is neither any indication as to the direction nor the amount of difference, nor what that quality is that is really being measured.

Analysis of variance.-- The difference between means can be evaluated only by considering also the standard deviation of the scores from their means. Therefore the variance is one of the two indications (the other being the mean) which must be examined if two or more groups are to be compared. If there are many groups, there are many pairs of means and variances to evaluate. This is a tedious process and further, there are many estimates of the real variance of the population, each estimate being derived

from one group sample. In addition, the real mean of the population is in question, although its stability can be estimated by the use of n-1 in the denominator of the formula

$$\sigma_M^2 = \frac{\sigma^2_{pop.}}{n-1}$$

where σ_M is the standard error of the mean, and n-1 replaces the usual n for all sizes of samples. This is an unbiased estimate of the standard error of the true mean for large as well as for small samples. The proof of this statement may be found in Lindquist [1].

Restating the concept of analysis of variance, it is possible to consider the variability of individuals about the mean of the group to which they belong, and "average" all those variations. The plan is to use many small groups, each from a different environment and each being tested under all sorts of conditions of personal attitude, health, and ability. This done, there should emerge a "normal level" of variability, "normal" for that population of which the groups are samples. This "normal level" of variability will now serve as a referent with which to compare the variance of any group (using the F-test) to see whether the group is like the general population. Stated as a null hypothesis, "The group is a member of the general population

---

[1] E.F.Lindquist, op. cit., p.50.

under consideration", and the hypothesis can be tested by means of the F-test, yielding an exact level of confidence at which we can accept or reject the hypothesis. Because the fluctuations of scores around the means of the small groups, whose members were chosen randomly, are partly due to errors of measurement and limited sampling, the average chance variance around the means of all groups is often called the "error variance". Obviously, if comparisons are made between the means and between the variances of any two groups, it may be mistakenly considered that this error variance is of the same quality and due to the same causes as the differences due to variations in the amount of possession of a trait or traits which the test purports to measure. In other words, there will oftentimes appear spuriously large variations which will be ascribed to differences in amounts of the trait present. Conversely, there may be present real and significant variations in the amount of the trait or traits which may be masked by larger chance variations. The analysis of variance therefore uses the error-variance of the whole population, as estimated from a composite of the variations of the individuals around the group means, as a basis for its comparison by means of the F-test ratio, with the variance of the means of the groups around the mean of the means (composite mean) of all the

groups. If this difference as indicated by the F-ratio is greater than can be expected by chance at a predetermined level, it indicates that there is at least one group among those being examined that is atypical, and is not in line with the "normal level" of variance of the population being examined. If such a group is discovered, further search by means of the t-test is indicated. After significantly different groups are isolated by means of the t-test, there still remains the problem of the reason for the lack of conformity by one or more groups to the general configuration of the population. The solution of this problem, while it may be furthered by the application of some other techniques, among which are the correlational methods, still must remain a logical rather than a statistical problem, and finally must rest for its solution on the thinking of an individual or a group of individuals. It is rarely possible for investigators to say with absolute confidence that there are present definite causal relationships, especially simple relationships, among measurements in the field of education.

Evaluation of the Picture-test.-- In Chapter III some criteria were stated for choice of items and for their continued membership in the test. Three further statistical criteria were applied to the 70 items which were

present throughout the testing of 271 randomly-selected pupils and for the 18 pupils of the Talent Group.

Item-difficulty, as determined by the critical ratio

$$C.R. = \frac{P_U - P_L}{\sigma_{P_U - P_L}}$$

where P_U is the proportion of the upper one-quarter (of the scores made by pupils) and P_L is the proportion of the lower one-quarter passing the item. The standard error of the difference is obtained by the formula

$$\sigma_{P_U - P_L} = \sqrt{\sigma^2_{P_U} + \sigma^2_{P_L}}$$

This was calculated for each item by the use of an abac devised by Mosier and McQuitty,[1] using 289 cases, 72 in each of the upper and lower quarters. Table 15 [2] page 80 has in column II the critical ratios of all 70 items. At the 5 per cent level of confidence, the t is 1.99 for 71 (72 in each quarter minus 1) degrees of freedom, and those items which could not meet this level were discarded in the final scoring of the test. Such items have their critical ratios marked with an (x), as have also a few items whose proportions were too near the 100 per cent or 0 per cent limits to be retained, as the critical

---

[1] Charles I. Mosier and John V. McQuitty, "Methods of Item Validation and Abacs for Item-test Correlation and Critical Ratio of Upper-Lower Difference", Psychometrika, Vol. 5, No.1, March, 1940, p. 57-65.

[2] Appendix B page 80.

ratio is not reliable at these extremes. A few non-discriminating items in the last four sub-tests were retained as clues. A further loss of the whole first sub-test, which was then designated "practice", occurred because of the use of a science achievement test different from that which was used on the other grades. With the items lost because of their non-discriminatory nature, and those of the practice area not scored, the test consisted, for scoring purposes, of 43 highly discriminatory items, varying in difficulty within each sub-test from very easy to very difficult. It should be noted that the item-analysis technique employed, while it was the best that could be discovered, probably caused a considerable loss of good items in a test of this type. The test is unique in having each item a cue-picture for all the other items of that sub-test. For that reason, the non-discriminatory items will be retained so that further investigation will yield results comparable with the statistics obtained thus far. A justification for using the scores from the discriminatory items after discarding the errors made on the non-discriminatory items is found in an article by Horst [1], who suggests not one but two such eliminations of error scores.

---

[1] Paul Horst, "Item Selection by Means of a Maximizing Function", Psychometrika, Vol.1, No. 4, December,1936, p. 229-244.

The item-test correlation was determined for each item by another abac given in the same article by Mosier and McQuitty [1]. Results will be found in Table 4 , page 50 78, column IV . The correlations were calculated for use in the estimation of the internal reliability of the test. No logical basis could be found for the splitting of the test into any pair of halves. The items are so closely related and there are so many sub-tests that any split-halves method is entirely unsuited to this study. Further, Kuder and Richardson have shown [2] that any split-half coefficient is only one of a multitude of such coefficients, depending on how the test is dichotomized. Of all other methods of deriving the coefficient of reliability, they conclude that the use of the item-test correlations of the items, together with the product pq, will yield the best coefficient of reliability, where p is the proportion of cases passing an item, and q (1-p) is the proportion failing the item, pq also being the variance of the item and a measure of its difficulty. In the same article they have determined formulas for various cases, each case assuming some facts about a test. Case II [3] fits the Picture-test most closely:

---

[1] Charles I. Mosier and John V. McQuitty, op.cit.,p. 60.

[2] G.F.Kuder and M.W.Richardson, "The Theory of the Estimation of Test Reliability," Psychometrika, Vol. 2, No.3, September, 1937, p. 151-164.

[3] ibid., p. 155.

> The data required are the number of items on the test, the difficulties of the items, the item-test correlations, and the standard deviation of the test. It is assumed that the matrix of inter-item correlation has a rank of one.

While it is impossible to check the assumption about the rank of inter-item correlations, and indeed the Picture-test may very likely not fulfill this condition, it is also pointed out that these inter-item correlations are but little more than 2 per cent effective (with 43 items) in changing the reliability of the test as a whole.

The formula used is:

$$r_{tt} = \frac{\sigma_t^2 - \Sigma pq}{2\,\sigma_t^2} \pm \sqrt{\frac{\Sigma r_{it}^2\, pq}{\sigma_t^2} + \left(\frac{\sigma_t^2 - \Sigma pq}{2\,\sigma_t^2}\right)^2}$$

where $\Sigma r_{it}^2 pq$ is the sum of the products of the squares of each of the item-test correlations by its pq product, and σ_t^2 is the variance of the whole test.[1] Only positive values of the radical are admissible in actual use.

Kuder and Richardson further state:[2]

> Reliabilities obtained from the formulas here are never overestimates. When the assumptions are rigidly fulfilled the figures obtained are the exact value of test reliability as herein defined; if the assumptions are not met, the figures obtained are underestimates....It is implicit in all formulations

---

[1] See data on test, Appendix B , p. 82..

[2] G.F.Kuder and M.W.Richardson, op.cit. p. 159.

> of the reliability problem that *reliability is the characteristic of a test possessed by virtue of the positive intercorrelations of the items comprising it*.

Using the formula above, the reliability of the test using the 43 discriminatory items was found to be

$$r_{tt} = 0.784$$

A measure of the stability of this r is the standard error of r. One formula for an unbiased estimate of this statistic is

$$\sigma_{r_{tt}} = \frac{1-r^2}{\sqrt{N-2}}$$

and

$$\sigma_{r_{tt}} = 0.0227$$

If this value is used to determine the critical ratio between the coefficient of correlation and its standard error, the quotient is 34.0. As a critical ratio of 3.0 is very significant, the coefficient of reliability of the Picture-test may be considered extremely stable.

A better statistic for the evaluation of the stability, upon repeated measurements, of the coefficient of self-correlation is found in the t-test. The formula;

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{N-2}$$

gives a value for t of 21.39. This t is extremely high, and would serve to indicate significance at the 1 per cent

level for even two degrees of freedom, instead of for the 269 d.f. here available. For information about the application of the formula above, and for the table of t-values used, the reader is referred to Lindquist.[1]

Another indication of the reliability of a test is discussed by Guilford [2]:

> Another idea of reliability conceives of the "true" scores for a group of individuals and asks how closely the obtained scores correlate with these true scores. True scores....are defined as those the testee would receive if the test were perfectly reliable....and this correlation is definitely related to the correlation between the forms (or halves) by the equation,
>
> $$r_{1\infty} = \sqrt{r_{11}}$$
>
> where $r_{1\infty}$=correlation between obtained scores and true scores .

That is, the "index of reliability" just defined above is the square root of the coefficient of reliability. For the Picture-test this index of reliability is

$$r_{1\infty} = \sqrt{0.784}$$

$$= 0.885$$

Returning to the coefficient of reliability, it is known [3] that this statistic indicates the proportion of

---

[1] E.F.Lindquist, op. cit., p.211, and Table 3, p.53.

[2] J.P.Guilford, Fundamental Statistics in Psychology and Education, McGraw-Hill Book Company, Inc., New York, 1942, p. 278.

[3] ibid., p. 223 and 278.

true variance present when "r" is squared. This statistic is the "coefficient of determination". The remaining proportion of variance $(1-r_{tt})$ is the error variance. In the Picture-test, therefore, 61.5 per cent of the variance is due to real differences in the possession of the trait or traits which the test is measuring, and 38.5 per cent is due to the error variance.

Distributions and Correlations.-- In this experiment, information about the Picture-test is sought through the analysis of the data gathered from a selected population. Three statistics were obtained for each pupil. The first is the score on the Picture-test, the second is the score on the achievement-test in general science, and the third is the I.Q. Some facts obtained follow.

1. Scores on the Picture-test were distributed for pupils in grades 8 through 12, not including the maximum-science group, as indicated in Figure 1 below.

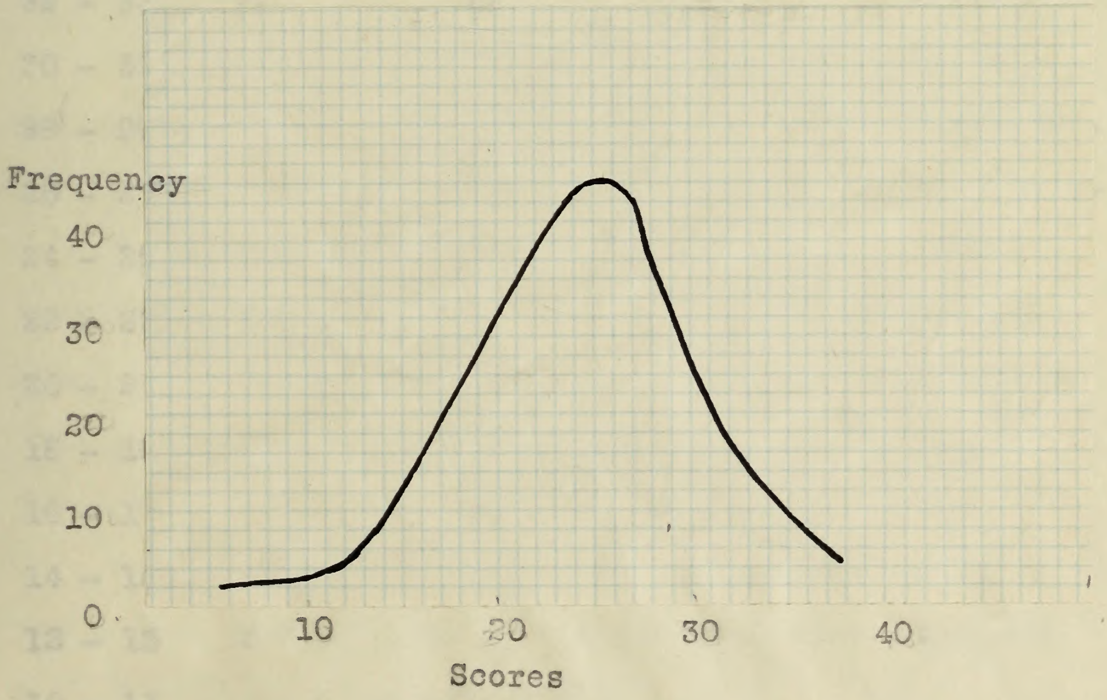


Figure 1. Smoothed Curve Representing the Distribution of 183 Scores Made by Pupils of Grades 8-12 on the Picture-Test. No Maximum-Science Pupils Included.

2. Some comparisons are made, in Table 1 below, among the various groups of pupils.

Table 1 . Comparison of the Distribution of Scores on the Picture-Test Made by Three Groups; and a Composite Distribution Including All Scores.

Scores	Grades 8-12, no maximum science	Maximum science	Talent Group	Composite of all scores
36 - 37	.	::	:	V:
34 - 35	:.	V:	:	VV:
32 - 33	V:	V:	::	VVV:.
30 - 31	V::	VVV	::.	VVVVV::
28 - 29	VVVVV.	VVV.	::	VVVVVVVVV.
26 - 27	VVVVVV	VVV	.	VVVVVVVVV.
24 - 25	VVVV::	V::		VVVVVV:.
22 - 23	VVVVV::	V.		VVVVVVV
20 - 21	VVVVV	V.		VVVVVV.
18 - 19	VV:.	.		VV::
16 - 17	V::	.		VV
14 - 15	::			::
12 - 13	:			:
10 - 11				
8 - 9	.	.		:
Totals	183	88	18	289

Key: V's are five scores; (:) are two; (.) is one.
Adapted from several sources.

Column (2), the basic non-maximum science group shows practically normal distribution, with a slight tendency toward negative skewness. The curve of Figure 1 confirms this normality. The maximum-science group, column (3) and the Talent Group, column (4) give the whole distribution as seen in column (5) a decided skewness, which is to be expected with the highly selected population which took the test.

3. Correlation of the scores on the Picture-test and the I.Q.'s, using the Pearson product-moment method, gives for $r_{xy} = 0.392$ and $\sigma_{r_{xy}} = 0.0514$

when $$r_{xy} = \frac{\frac{\sum XY}{N} - M_x M_y}{\sigma_x \sigma_y}$$

A better estimate of the reliability of r is the t-ratio, [1] $t = 6.99$

when $$t = \frac{r}{\sqrt{1-r^2}} \sqrt{N-2}$$

This t is so large as to discredit the hypothesis that there is no correlation between the two measures.

A scatter diagram of 170 randomly-chosen original raw scores against I.Q.'s is shown in Figure 2 ,page 46 .

---

[1] E.F.Lindquist, op.cit. p.211.

While there is evidence of linearity of regression, there are some anomalous placements of cases. The few scores near the area (A) are probably true non-readers, able to do well on the test because of its non-verbal nature although high scores on a written group intelligence test are impossible. A more important aspect of the diagram is the indication that while high I.Q.'s are necessary for the making of high scores on the Picture-test, they do not insure such scores, as seen in area (B). This is an example of the old saying that "A man may fall down, but never up", which applies to mental as well as to physical achievement.

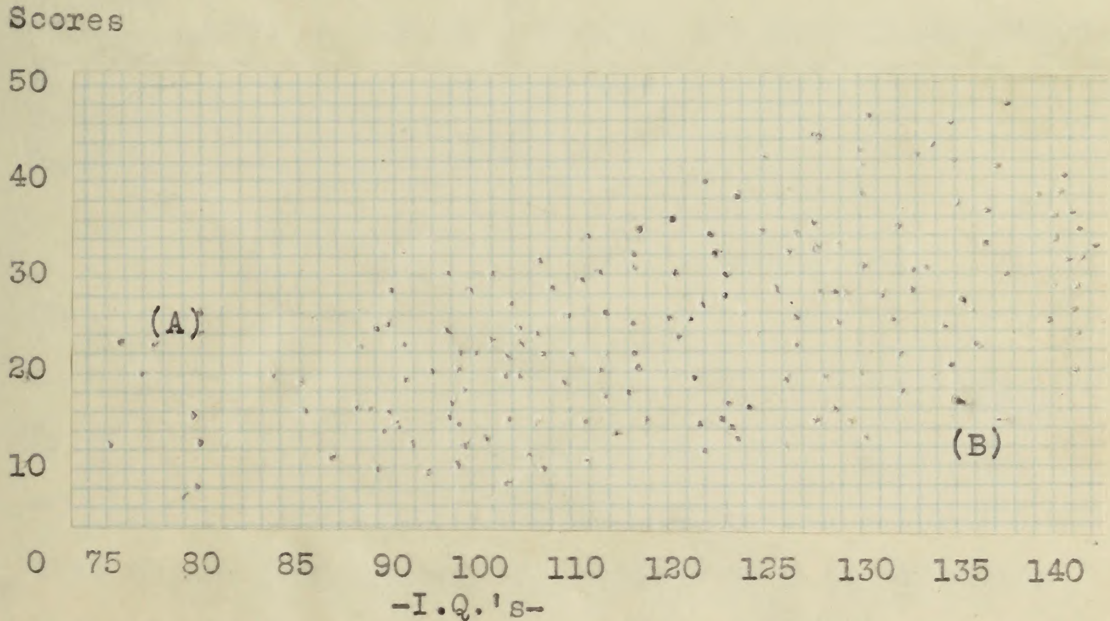


Figure 2. Scatter Diagram of 170 Randomly-Selected Original Raw Scores on the Picture-Test Plotted Against I.Q.'s.

4. The data on the degree of correlation between the scores on the Picture-test and those on the general science achievement-test, using the same formulas as those used above, gave

$$r = 0.533$$

$$\sigma_r = 0.044$$

$$t = 10.33$$

Again, t is so large as to discredit the null hypothesis that there is no relationship between scores on the Picture-test and scores on the achievement-test.

Data from three tests.-- Three measuring devices were used in the experiment. In Table 2 below are figures derived from scores on the Picture-test, from I.Q.'s, and from scaled scores on the science achievement-test.

In the paragraphs that follow, numbers of pupils vary. Scores were withdrawn randomly to make groups of sizes convenient for analysis.

Table 2 . Data from Three Tests Given to Pupils Participating in the Experiment.

	N	Mean	σ_s^2	σ_p^2	σ_s	σ_M
Picture-test....	271	25.36	26.27	26.37	5.13	0.312
I.Q.......	271	108.98	214.18	214.98	14.64	0.830
Science test....	271	49.32	174.70	175.34	13.21	0.803

Key for all tables: N or n, number of cases; σ, standard deviation or standard error; s, sample; p, population; M, mean.

Normality of Picture-test score distribution.-- The application of the F and t statistics assumes randomly-selected small samples from distributions which are nearly normal. Using the goodness-of-fit application of chi-square,[1] the correspondence of the observed frequencies with theoretical normal frequencies of a distribution whose mean and standard deviation were the same as those of the Picture-test was examined.

---

[1] E.F.Lindquist, op. cit., page 37.

The observed chi-square was 6.908. This value falls between the 70 per cent and the 50 per cent levels of confidence. There is no reason to reject the hypothesis of normality.

Examining the hypotheses of the design.--The four hypotheses set up at the beginning of Chapter IV can now be examined.

1. The first hypothesis predicated no difference between mean scores made by boys and mean scores made by girls. Data are given below in Table 3.

Table 3 . Comparison of Scores Made by Boys and Those Made by Girls on the Picture-test.

	N	Mean	σ_s^2	σ_p^2	σ_s	σ_M
Girls....	126	23.77	24.70	24.89	4.97	0.445
Boys.....	138	26.74	23.92	24.00	4.88	0.415

The F-ratio is 24.89/ 24.00, and is equal to 1.04. This is not significant at the 5 per cent level (where 1.34 is required). We cannot state with much confidence that there is any real difference between boys and girls in performance on the Picture-test.

2.The second hypothesis stated that there are no differences among the means of scores obtained by pupils of various grade levels 8-12. Table 4 ,page50 ,gives the information about this hypothesis.

(In the table below, the groups have been reduced to six or twelve pupils each for use in the analysis-of-variance to be carried out presently. A total of 220 pupils includes representative samples of all groups.)

Table 4 . Grade Level Achievement on the Picture-test.

Grade	n	Mean	σ^2_s	σ^2_p	σ_s	σ_M
8......	48	22.50	23.46	23.96	4.85	0.707
9......	48	24.12	22.81	23.30	4.76	0.694
10......	24	25.62	21.70	22.64	4.67	0.953
11......	66	26.95	22.17	22.50	4.71	0.579
12......	54	26.33	28.31	28.84	5.32	0.731

It is evident that the largest F-ratio is that between grades 11 and 12. The population variances yield for 28.84/ 22.50 an F of 1.28. For the 5 per cent level of confidence an F of 1.54 is required. There is no reason to discredit the hypothesis that there is no difference in performance on the test due to grade-level.

3. The third hypothesis was that there is no difference between means of scores made by pupils in the same grade but from urban or suburban areas. Grades 8 through 10 are described in Table 5 , page51 , reserving grades 11 and 12 for the more complete analysis- of-variance technique. It should be recalled here that urban areas are those with more than 25,000 population, while suburban places have populations under 7,000.

Table 5. Comparison of Scores Made by Pupils of Grades 8,9, and 10 in Urban and Suburban Schools.

Type of School	n	Mean	σ^2_s	σ^2_p	σ_s	σ_M
Grade 8						
Urban...	24	23.29	19.57	20.43	4.42	0.992
Suburban	24	21.71	26.08	27.22	5.11	1.07
Grade 9						
Urban...	24	25.50	23.41	24.43	4.84	1.01
Suburban	24	22.75	18.19	18.99	4.26	0.888
Grade 10						
Urban...	12	23.17	18.56	20.25	4.31	1.24
Suburban	12	28.08	12.50	13.64	3.53	1.02

F-ratios comparing variances of urban and suburban schools are: in Grade 8, 1.33; in Grade 9, 1.29; in Grade 10, 1.48. Inasmuch as an F of 1.98 is needed for significance at the 5 per cent level for the larger degrees of freedom (23 and 23) and more than that figure for the smaller degrees of freedom, (11 and 11), there is no reason to discredit the hypothesis that there is no difference between the performances on the test of the urban and the suburban groups.

4. The final hypothesis, that the maximum-science groups are like the minimum-science groups in their abilities to score high on the test, involves the analysis of variance. Table 6, page 52, shows the method of calculation used.

Boston University
School of Education
Library

Table 6 . Analysis of Variance in Grades 11 and 12, Maximum and Minimum Science Groups.

N	120	all cases
n	6	number of cases in each group
G.T.	3201	Grand Total of all scores
G.M.	26.675	General Mean of all scores
$(\Sigma x)^2$	518,629	Find the sum of the scores of each group; square each of these sums; sum all the squares.
(Σx^2)	88,377.00	Square each individual score; sum all the squares.
GT x GM	85,386.68	Multiply the G.T. by the G.M.
Sum of Squares (Total)	2,990.3	Find the difference between last two quantities.
Sum of Squares (Max. and Min. Science)	1,051.5	Divide $(\Sigma x)^2$ by n; then subtract GT x GM.

Sum of Squares		d.f.	Variances
(Max. and Min. Science)	1,051.5	19	55.34
Within Groups (Error Variance)	1,938.8 1/	100	19.39
Total..........	2,990.3	119	25.13

1/ This figure is obtained as the difference of the Total and the Sum of Squares (Max. and Min. Science).

The "Sum of Squares (Max. and Min. Science)" is here the "between groups" variance calculation. In this table there are 20 such groups of 6 pupils each, among which groups differences are being sought. The 20 groups result in 19 degrees of freedom; the Total has one d.f. less than the N of the array. As suggested by the footnote, the "Sum of Squares " of Within Groups is found by subtraction, and so is the figure for d.f. for that category, 100. The variances are quotients resulting from the division of each "Sum of Squares" by its corresponding d.f. The whole system of calculation, together with a discussion of the limiting factors and the interpretation of results, may be found in Lindquist. 1/

A significant F is seen to be present if the variance of "between groups" is divided by the variance of "within groups", or 55.34/ 19.39. This F is 2.85 and is significant. The null hypothesis postulated at the beginning of this problem, that there are no differences among the groups, is therefore discredited at the 1 per cent level. Differences among the means of the groups may be sought.

One possibility is that the means of the scores of boys and girls in the minimum-science group are alike. Table 7 , page 54 gives the data on this group.

1/ E.F.Lindquist, op.cit., p. 100-104.

Table 7 . Comparison of Scores Made by Minimum-Science Boys and Girls.

Sex	n	Mean	σ^2_s	σ^2_p	σ_s	σ_M
Boys.....	24	26.46	14.95	15.60	3.86	0.804
Girls....	24	23.71	19.41	20.26	4.40	0.917

The *F* is 1.30, and is not significant. The hypothesis cannot be overthrown at the 5 per cent level.

Another hypothesis may be stated;that boys and girls of the maximum-science group are alike. Table8 below presents this case.

Table 8 . Comparison of Scores Made by Maximum-Science Boys and Girls.

Sex	n	Mean	σ^2_s	σ^2_p	σ_s	σ_M
Boys.....	42	29.54	13.76	14.09	3.71	0.577
Girls....	30	25.20	32.36	33.48	5.69	1.040

The *F* here is very significant, as it is 2.38 when 2.17 only is needed at the 1 per cent level. The formula for *t* when the two samples are not members of the same population (see p. 30) gives for the comparison of the means a value for *t* of 3.19. The largest *t* listed for 1 per cent significance is 2.58. It is possible to reject the hypothesis that maximum-science boys and girls are alike in their ability to score on the Picture-test.

Another possibility is that urban and suburban pupils in grades 11 and 12 are alike. Table 9 below presents data for the evaluation of this hypothesis.

Table 9 . Comparison of Scores Made by Urban and Suburban Groups in Grades 11 and 12.

Place	n	Mean	σ_s^2	σ_p^2	σ_s	σ_M
Suburban..	72	27.70	18.85	18.97	4.34	0.515
Urban.....	48	25.19	12.33	12.59	3.49	0.509

The F is not significant at the 5 per cent level. However, it may be noted that an F of 1.57 is significant at that level. We cannot discredit the hypothesis at the 5 per cent level, but note that there may be some significance here, for the F is 1.51.

The data of the next table apply to the fourth hypothesis of the original design of the experiment, that there is no difference between mean scores of those who have had maximum-science and those who have had minimum-science training. It would seem that this might have been decided directly without any recourse to the analysis-of-variance techniques. Indeed it might have been, though this would have left unanswered all the questions that have thus far been posed about sub-groups in this area. Table 10 ,page 56 completes the study of the four hypotheses of the experimental design.

Table 10. Comparison of Scores Made by Minimum-Science Pupils and Those Made by Maximum-Science Pupils in Grades 11 and 12.

Group	n	Mean	σ_s^2	σ_p^2	σ_s	σ_M
Maximum-Science.	72	25.73	26.14	26.50	5.11	0.614
Minimum-Science.	48	25.08	19.20	19.61	4.38	0.632

The *F* at the 5 per cent level for the d.f. above is 1.57, and the *F* reached in the comparison of the maximum and minimum-science groups was only 1.35. The hypothesis that the two groups are alike cannot be discredited.

Finally, the Talent Group is compared below in Table 11 with the maximum-science group to see if they are alike.

Table 11. Comparison of the Talent Group with the Maximum-Science Group.

Group	n	Mean	σ_s^2	σ_p^2	σ_s	σ_M
Talent Group...	18	31.44	7.49	7.90	2.74	0.646
Maximum-Science.	72	27.73	26.14	26.50	5.11	0.602

The *F* of 3.35 is very significant . The t-test, applied to the difference between the means (page 30) gives a *t* of 2.94. This is significant at the 1 per cent level. The hypothesis that the two groups are alike may be rejected.

CHAPTER VI.

CONCLUSIONS

Reliability of measurement.-- The Picture-test measures some trait or traits consistently. At the 1 per cent level it may be said that the variability, as measured by the variance, is the same for all grades 8 through 12, for boys and girls, and for urban and suburban groups. There is a small but steady increase in the means of grades 8 through 11, and the slight loss in grade 12 is probably due to sampling error. These mean scores do not indicate that the number of years of school attendance is a factor that can be said to affect scores on the test at the 5 per cent level. Neither do the number of years of science affect the means of the scores, as is evidenced by the failure to discover a difference (again at the 5 per cent level) between means of the total maximum science group and means of the total minimum-science groups. However, at the 1 per cent level the boys of the maximum-science group can be said to be superior to the girls of this group.

The criterion group, called here the Talent Group, exhibited a very significant superiority at the 1 per cent level. This group, with a mean of 31.44 and an estimated

population variance of 7.90, is seen to be outstanding when it is compared with the maximum-science group whose mean is 27.73 and whose estimated population variance is 26.50. The process of selection that resulted in membership in the Talent Group strongly evidenced itself in their scores on the Picture-test. This group also offers evidence, on a small scale, that numbers of years of school or of science are not the deciding factors in the obtaining of high scores on the test, for six of the eighteen members of the Talent Group were from grades 7, 8, and 9 and had had neither physics nor chemistry. It may be that the Picture-test is measuring a different kind of trait or of "intelligence" than is measured by tests currently in use.

The correlation of the Picture-test scores with I.Q.'s was 0.392. This is markedly lower than the correlation usually reported between intelligence test-scores and achievement test-scores. With the general science achievement-test, the raw coefficient of correlation was markedly higher, 0.533, and probably indicates that the Picture-test is measuring at a significant level many of the same factors which the science test measures.

The Picture-test is admittedly short, and sub-tests in other areas could be devised. For the whole test, the coefficient of reliability is 0.784. The index of reliability, which is the maximum coefficient of reliability

which the test could have in its present form and length, is 0.885. The test could profitably be lengthened to 120 items. Test-builders may be interested to note that this was close to the original number of items in the Picture-test, so that if further experiments with this type of test are undertaken, one-half to two-thirds of the original items may be expected to be lost in the item-analysis.

It is possible that variable factors which were uncontrolled in all the three tests which were used to gather data from the pupils of the experiment,made it impossible for correlations among them to be as high as they might have been. Correction for attenuation in both measures of a pair would raise the coefficients considerably, or if only the criterion-tests (the intelligence and science tests) were so corrected, would raise the coefficients to a somewhat smaller degree. Partialling-out intelligence or science achievement would not usually give a coefficient higher than the highest member coefficient. However, the randomization practiced in the experiment can almost certainly be said to have cancelled out many of the variable factors under discussion. The sample of the study, chosen randomly as it was, should be free from bias, and should yield statistics which are as reliable as those which might be obtained from a huge unrandomized sample from the same schools.

Validity indications.-- As there was no single criterion against which this test could be evaluated, the validity of the test is measured largely by pupil-responses to it. This situation was recognized at the outset of the experiment, and the study was designed to furnish some information about the test.

If then a test exhibits properties like those discussed above, and discovers that high scores on the test are made by pupils without regard to many of the factors associated with school populations, there are two possibilities. The first is that the scores are distributed solely by chance, and that the test is therefore useless. This hypothesis is untenable on two counts. One of them discredits the hypothesis at the 1 per cent level, for the correlations with both the other tests are of such sizes that they could not (at better than the 1 per cent level of confidence) have arisen by chance, as is shown by their t statistics. The other count makes it clear that the definite superiority and the extremely low variance of the Talent Group combine to insist that their test scores are the result of the possession of some trait or traits which the test is measuring. This Talent Group, made up as it was of pupils from all grades 7 through 12, is a further reason for believing that the Picture-test finds high scores where they are, and that little more can be said about places in which

to look for such pupils. There are a few statistical pointers which might indicate superiority where none is indicated at the 5 per cent level. F-ratios indicate that suburban pupils may sometimes be better than urban pupils, and maximum-science pupils may sometimes be superior to minimum-science pupils.

Uses for the Picture-test.-- Young people in our schools who are talented in science are being sought by colleges and industrial organizations. The Picture-test in expanded form and with many more areas sampled in its sub-tests should supply information about those young people, which is not now obtained from the administration of either the written intelligence-test or from the science-type written achievement-tests. The Picture-test might become one of the increasingly varied battery of measures which yield information about pupils so as to predict their future success.

The test may be used, as it is, to help select candidates for the school science club or for scholarship awards in science.

Wherever a situation arises where it is helpful to know the individual's ability to recognize and analyze science-type problems, there is a strong possibility that the Picture-test will give unique and dependable information. If it is agreed that the scientific method of approach to problems is also the method of critical thinking, then this test

may be used to determine the level at which such thinking is done, in other than science areas.

CHAPTER VII

SUGGESTIONS FOR FURTHER STUDY

Prediction possibilities.-- Pupils in junior high school often show interest and ability in science. The Picture-test may be able to predict future success for these pupils in the more formal sciences of high school and college, and so fix in their minds the idea that they perhaps are potentially science majors. A great deal of care should be exercised here, as continuance with science also presupposes decided excellence in mathematics and high general intelligence. Science today is no longer a profession that calls only for ingenuity and for problem-solving capacity, although these are necessary. Science today is based almost wholly on mathematics at a very high level. While most able science students do go on with their studies, the saving of just one or two students annually for science may be important. The individual contributions of workers in the field of scientific research are sometimes important far beyond their quantitative value. As of today (1946) there is, and will be for a decade, a serious shortage of scientists at a time when science problems are increasing exponentially. A ten year study

is indicated to follow seventh-grade pupils into college.

The problem of elementary or secondary school science education is largely the problem of finding adequately trained and able teachers. Again, the Picture-test purports to test but one of the factors that is necessary for complete understanding of the purposes of science and science education. Either for entrance upon a teaching career in science, or indeed for entrance to a school of education, the Picture-test might furnish some evidence of the student's problem-solving ability. While the test is not sufficiently validated at other levels than those where it was tried, both the test and the testees would be experimental material. It is likely that the only way in which the test will achieve greater usefulness will be by controlled experiments like the one just completed.

Teaching with the test.-- Not only can the device used here for picture-matching be used for testing, but it may be used as a teaching-device that will at least call attention to the nature of problem-solving activities. Subject-matter is relatively easy to teach. Most teachers dealing with science-type material, together with the texts they use, profess to teach the scientific method. Most of them also would admit that they do very little such teaching, or if they claim so to teach, can show objective

evidence of increments in the ability to solve problems by the method of critical thinking. And so many continue to teach a great deal of subject-matter, not only because it is easy to teach, but because it is easy to test results. Perhaps the Picture-test can provide material and method for the teaching of facts and scientific method, and then, having adequately pre-tested the ability to solve problems, measure small increments in that ability.

The Picture-test a prototype.-- Most important of all, the type of test may be applied to a variety of fields. All science-type subjects should find the form of the test valuable. Biological science, geography, health and medicine, safety, and others have real situations which might be photographed to provide innumerable items. In vocational and educational guidance, in the area of citizenship and ethics, and in the region of personal-problem solving, where relationships with other people are concerned, there should be fruitful research possible. If the study now completed has made a discovery, it is that another tool for the evaluation of human abilities has at least been found to hold an edge and to cut a small shaving cleanly.

BIBLIOGRAPHY

Barry, Robert, "An Analysis of Some New Statistical Methods of Selecting Test Items", Journal of Experimental Education, Vol. 7, 1938-39, p.221.

Beauchamp, Robert O., and Hanor A. Webb, "Resourcefulness, an Unmeasured Ability", School Science and Mathematics, Vol. XXVII, May, 1927, p. 465.

Bennett, George K., Test of Mechanical Comprehension, Form AA, The Psychological Corporation, 522 Fifth Ave., N.Y., 1940.

Bode, Boyd Henry, Conflicting Psychologies of Learning, D.C.Heath and Company, Boston, 1929, 300 p.

Brody, Leon, "Comparable Tests of Verbal and Non-verbal Reasoning", Journal of Educational Psychology, Volume 31, No. 4, April, 1940, p.180-194.

Burt, Cyril, The Factors of the Mind, The Macmillan Company, New York, 1941, 509 p.

Cattell, Raymond B., "A Culture-free Intelligence Test,I", Journal of Educational Psychology,Vol. XXXI, No.3, March 1940, p. 161-179.

Curtis, Francis D., Second Digest of Investigations in the Teaching of Science (1925 - 1930), P. Blakiston's Sons, and Company, Inc., Philadelphia, Pa., 1931, 424 p.

________________, A Synthesis and Evaluation of Subject Matter Topics in General Science, Ginn and Company, Boston, 1929, 83 p.

Dodd, Stuart, The International Group Mental Test, Doctorate Thesis, Princeton University, 1926.

Durrell, Donald D., and Helen Blair Sullivan, "Manual for Primary and Intermediate Tests, Durrell-Sullivan Capacity and Achievement Tests", World Book Company, Yonkers-on-Hudson, New York, 1937.

Edds, Jess H., "The Nature of Verbal and Non-Verbal Abilities", Journal of Experimental Education, Vol. III, No. 3, 1935, p. 225-229.

Fisher, R.A., and F. Yates, Statistical Tables for Biological, Agricultural and Medical Research, Oliver and Boyd, London, 1938, 90 p.

Fisher, R.A., Statistical Methods for Research Workers, Oliver and Boyd, Edinburgh, 1925, 344 p.

Goldstein, Kurt, and Martin Scheerer, "Abstract and Concrete Behavior", Psychological Monographs, Volume 53, 1941, p. 1-149.

Guilford, J.P., Fundamental Statistics in Psychology and Education, McGraw-Hill Book Company, Inc., New York, 1942, 333 p.

Halstead, W. C., "Preliminary Analysis of Grouping Behavior in Patients with Cerebral Injuries", American Journal of Psychiatry, Volume 96, 1940, p. 1-64.

Healy, W., "A Pictorial Completion Test", Psychological Review, Volume 21, 1914, p. 189-203.

Horst,Paul, "Item Selection by Means of a Maximizing Function", Psychometrika, Vol.1, No.4, December, 1936. p. 229-244.

Keeslar, Oreon, "The Elements of Scientific Method", Science Education, Vol. 29, No. 5, December 1945,p.275.

Korzybski, Alfred, Science and Sanity, The International non-Aristotelian Library,Publishing Co., The Science Press Printing Company, Scranton, Pa., 1941, 806 p.

Kuder, G.F., and M.W. Richardson, "The Theory of the Estimation of Test Reliability", Psychometrika, Vol. 2, No. 3, September, 1937, p. 151-164.

Lindquist, E. F., Statistical Analysis in Educational Research, Houghton Mifflin Company, Boston, 1940, 266 p.

Mosier, Charles I., and John V. Mcquilty, "Methods of Item Validation and Abacs for Item-Test Correlation and Critical Ratio of Upper-Lower Difference", Psychometrika, Vol. 5, No. 1, March, 1940, p. 57-65.

National Society for the Study of Education, "A Program for Teaching Science", Part I, The Thirty-First Yearbook. Public School Publishing Company, Bloomington, Illinois, 1932, 370 p.

Noll, Victor H., Teaching of Science in Elementary and Secondary Schools, Longmans, Green and Company, Inc., New York, 1939, 238 p.

Powers, S. R., "Objective Measurement in General Science", Teachers College Record, Vol. 29, p. 345-349.

____________, "The Correlation between Measures of Mental Ability and Measures of Achievement in Chemistry", School Science and Mathematics, Vol. 28, December, 1928, p. 981-986.

Pritchard, John W., "Motor Performance as a Chance Factor in Test Scores", Journal of Educational Research, Vol. 37, No. 3, p. 181-192.

Searle, Albert H., and Giles M. Ruch, "A Study of Science Articles in Magazines", School Science and Mathematics, Vol. 26, April, 1926, p. 389-396.

Snedecor, G. W., Statistical Methods Applied to Experiments in Agriculture and Biology, Collegiate Press, Inc., Ames, Iowa, 1937, 422 p.

Sorenson, H., Statistics for Students of Psychology and Education, McGraw-Hill Book Co.,Inc., New York, 1936, 373 p.

Smith, Victor C., "A Study of the Degree of Relationship Existing between Ability to Recall and Two Measures of Ability to Reason", Science Education, Volume 30, No. 2, March, 1946, p. 88,89.

Sullivan, Helen Blair, "A New Means of Appraising the Qualifications of Prospective Teachers", Harvard University, Graduate School of Education, unpublished thesis, 1944.

Underhill, O. E., and S. R. Powers, Cooperative General Science Test, American Council on Education, Cooperative Test Service, New York City, 1939.

Weber, Edmund G., "Equating High-School Intelligence Quotients with College Aptitude Test Scores", The Journal of Educational Psychology, Vol. XXXVI, No.7, October, 1945, p. 443-446.

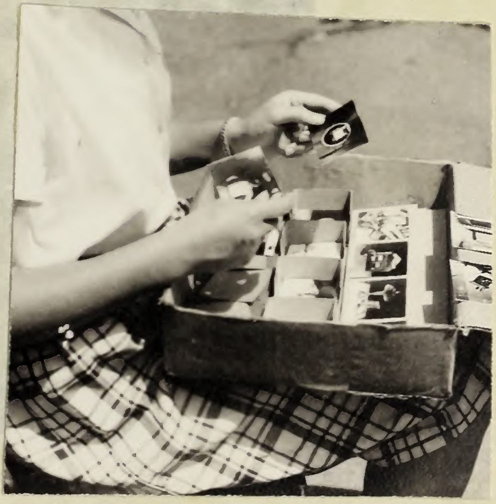
Wise, Harold E., "A Determination of the Relative Importance of Principles of Physical Science for General Education", Science Education, Vol. 25, No.7, December, 1941, p.371-379.

Appendix A

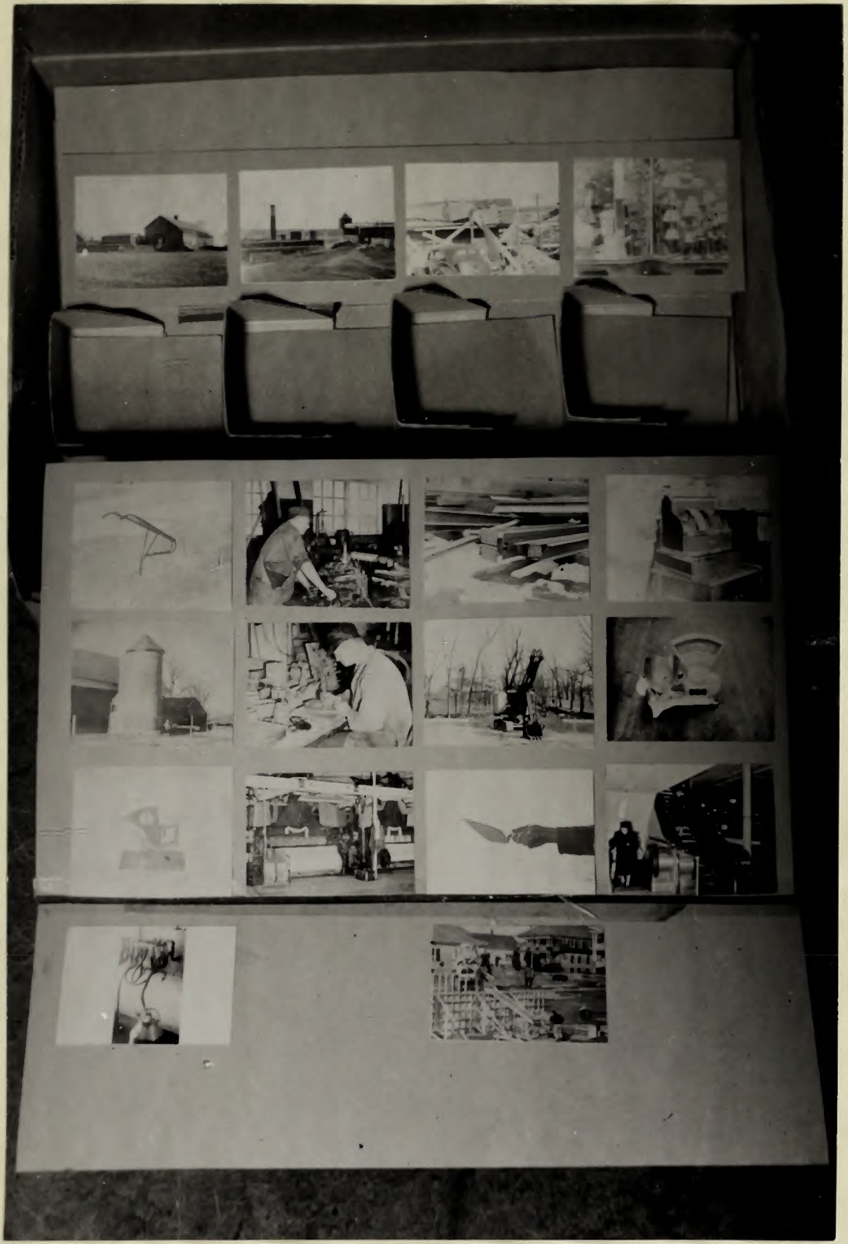
Giving Directions

The Test is Opened

Puzzling Out Cue-
Pictures



Delayed Judgment



Practice Area Items and Cue-Pictures

Appendix B

DIRECTIONS for taking ENVIRONMENT-CONSCIOUSNESS TEST

You have a list of words; you have also eight sheets with numbers from one to fifty on them. Write the title for each, taking the titles from the list below (one title on each blank list).

1. Astronomy
2. Aviation
3. Biology
4. Chemistry
5. Geology
6. Health
7. Physics
8. Weather

(Biology does not include human functions, disease, accidental injuries; these come under Health)

a. Now take the list of words and the blank sheet titled Astronomy. Read the list of words, starting with PART I. Find a word which belongs on the Astronomy sheet. Write it down. Go through all the whole list of words until you have 50 words. If you do not find 50 words that you recognize, put away the Astronomy list. (You have read Part I and Part II. of the word-lists)

b. Now take the sheet titled Aviation and do the same thing.
c. Do this with all eight sheets.

d. Occasionally a word may appear in two categories; that is, you may find that a. word may be used twice. It is all right so to use the words.

e. Every word under the correct heading will count one point. No credit will be lost if incorrect words are placed on any sheet.

f. Try to get 50 words on each list, but you must do all lists. You will need to work fast, as there are 400 words to read eight times !

g. MAKE NO MARKS OF ANY KIND ON THE PRINTED LISTS OF WORDS! RETURN THESE LISTS WITH YOUR COMPLETED TEST.

PART I- USE THIS LIST FIRST

lymph
helminth
ground crew
star
amanita
acceleration
reflux
exfoliation
barometer
rivet
~~luna~~
cellophane
polar front
plankton
hypertension
monadnock
blue earth
anvil
grid
hookworm
Scheele
~~tube~~
siderite
granite
maritime
lactation
fumarole
duralumin
fuse
orthopedics
flux-gate
saprophyte
cauterization
ether
alcohol
rachitic
tributary
archegonium
fairing
cumulus
tide
sunspot
vulcanism
gneiss
cambium
mesa
feathering
plastic
nitrogen
frost
appendectomy
tension
declination
sleet
~~lunar~~
cantilever

cotyledon
assay
dessicator
convection
light-year
plexiglas
cold front
refractory
earthquake
agriculture
henry
cat-paws
tsunami
ground-loop
Polaris
~~cemented~~
artificial horizon
malocclusion
celestial
benzene ring
tsuga
joule
equatorial mount
modulation
fix
Bernoulli
universe
air mass
pressure point
helicopter
locust
fermentation
resin
pediatrics
oscillations
millibar
insolation
nymph
trailing edge
precession
protozoa
saponification
dike
oxbow
Eridanus
ornithopter
beri-beri
peneplain
planetesimal
mitosis
thermodynamics
nephoscope
Kjeldahl
hygrometer
Aurora Borealis
katydid

geode
binary
villi
high
physiography
thermal
crystal
drizzle
parachute
parsec
intercom
Orion
equinox
pomology
oxidation
gale
aberration
Renaud
gilbert
deliquescent
rhus

sun
carburetor
insectivorous
pumice
seismograph
yaw
inclinometer
phenophthalein
splint
igneous
photosynthesis
cyanamide
style
sidereal
polystyrene
vitamin
glaciation
stalactite
novae
halo
polarization
delta
jet propulsion
catalyst
tropopause
umbra
chaulmoogra
laccolith
undercarriage
tappet
smuts
cirro-stratus
condenser
moon
carborundum

hibernation
mike
chromosphere
Venus
oersted
stratosphere
[illegible]
fossil
camber
infra-red
radiosonde
topaz
adiabatic
capacitance
gust
barium
canyon
~~midrib~~
bovine
meningitis
counterpoise
thunderstorm
dolphin
connecting-rod
archeopteryx
platypus
plasma
schizomycetes
nodule
mass
telescope
geyser
fathometer
~~sponsons~~
monsoon
kilowatt
diastrophism
parallax
suture
theodolite
sun-dog
ascorbic
ester
dune
Sirius
Pegasus
isomerism
lens
manifold
nucleolus
therapy
~~flaps~~
penicillin
centrifugal force

PART II.

planet
sound
sweepback
loess
compound
hurricane
Beaufort
lichen
Bangs
horsepower
vaccine
elevator
feldspar
solstice
malaria
sedimentary
~~[illegible]~~
cyclone
graft~~ing~~
winds
inertia
ionization
plant
mushroom
paraffin
solution
occultation
navigator
anode
pinus
comet
telephoto
DDT
hydroxyl
doldrums
arterial
lava
dead reckoning
chlorination
valence
caries
eruption
hour-angle
calorie
electron
kettle-hole
Mt. Palomar
chlorophyll
saccharimeter
stability
blizzard
asepsis
neap
arthritis
neuron

antitoxin
salt
ovule
nebula
brucella
oleo strut
lever
zenith
yaws
mineralogy
limestone
metamorphosis
cockpit
focus
sandstone
tornado
expansion
typhoid
botulism
aileron
animal
typhoon
humidity
asteroid
fireworks
cowling
anemometer
propeller
medulla
canals
pothole
conservation
solvent
feline
altimeter
isotope
gabbro
lumen
larva
pancreatic
dengue
resonance
distillation
sirocco
insulin
astigmatism
sextant
Lister
antitrade
polymerization
rain
Bohr
decibel
pilot

lime
waterspout
nylon
zodiac
fluid
piston
shale
supercharger
paleontology
thermograph
meter
vertebrate
typhus
hydraulics
spar
cyclopropane
coronary thrombosis
water-vapor
pistil
amphibian
antivenin
fog
constellation
pasteurized
hornblende
drag
elasticity
aspirin
B. coli
modulation
lift
isogonic
deuterium
eucalyptus
poundal
rapids
coherer
anemia
Archimedes
pyrethrum
equator
gibbous
till
foehn
conglomerate
Tesla
Corsair
vortex
radar
aldehyde
ion
butane

bolide
refraction
stick
bank
radial
galaxy
pneumonia
plateau
vacuum
base
friction
parasite
insect
algae
dew-point
watt
harmonic
squall
monococque
precipitation
hydrogenation
eclipse
erosion
isotherm
elephantiasis
garnet
TNT
sulfathiozole
vector
anesthesia
polyphase
radical
psychrometer
Capricorn
detergent
empennage
glider
Wright
intravenous
sequoia
bakelite
meander
cube
crucible
kame
endothermal
hydrolysis
hygrometer
Cassegrain
Foucault
Heaviside
tantalum

Table 12 . Correct Placement of Item-Pictures in Bins.

Bin 5	Bin 6	Bin 7	Bin 8
704	703	702	701
		706	705
605	602	604	601
606	603		
502	504	503	501
506	505		
402	401	403	404
	406	405	
315	307	306	305
318	311	309	308
320	313	310	312
322	319	314	316
	321		317
209	207	206	205
213	211	210	208
216	215	214	
218	217		
106	105	107	113
110	108	111	116
112	109	115	118
114		117	

Table 13. Item Numbers and Descriptions.

105	blanket loom	205	man being X-rayed
106	wheel cultivator	206	electric flatiron
107	pouring concrete	207	loudspeaker
108	metal lathe	208	X-ray machine
109	bench worker	209	elevator
110	silo	210	electric soldering-iron
111	gasoline shovel	211	radio broadcasting antenna
112	milking machine	212	telephone lineman
113	grocery scales	213	automobile starter-pedal
114	egg scale	214	exposure meter
115	mason's trowel	215	door bells
116	escalator	216	electric fan
117	girders	217	telephone switchboard
118	cash register	218	vacuum cleaner

Table 13, continued. Item Numbers and Descriptions.

305	oil refinery
306	voltmeter
307	dinosaur
308	burettes
309	oscilloscope
310	calipers and tape
311	sea-beach
312	chemical balance
313	deep-well drilling
314	stop watch
315	telescope
316	beakers
317	rolls in oven
318	longitude lines
319	air-view of glacier
320	sextant
321	fossil
322	meteor
401	fountain pen
402	air-drill
403	evaporator on ice-plant
404	airplane
405	CO_2 extinguisher
406	plumber's force-cup
501	arc-light
502	electric meters
503	lightning-rod
504	automobile headlight
505	battery hydrometer
506	transformer
601	gasometer
602	soda fountain
603	yeast-cakes
604	soap-bubbles in pan
605	oil heater
606	oil on gun
701	hand opening door
702	screw-top jar
703	spring and weight
704	child in swing
705	arm holding weight
706	bench vise

Table 14. Distribution of Pupils Participating in the Study.

School and Type	Grade	Boys	Girls	Maximum Science	Minimum Science
Urban					
A	11	12			12
A	12		12		12
A	12		12	12	
A	12			12	
Urban					
B	9	14	13		27
Suburban					
C	8	12	12		24
D	12	6	6	12	
D	12	6	6		12
Urban					
E	8	12	12		24
Suburban					
F	9	12	12		24
F	10	6	6		12
F	11	6		6	
F	11	6	6		12
F	12	6	6	12	
Urban					
G	10	6	6		12
G	11	6	6	12	
G	11	6	6		12
Suburban					
H	11	8	2	10	
Suburban					
I	11	6	6	12	
Talent Group	7-12	13	5	18	
Totals		155	134	106	183

Grand Total - 289

Table 15 . Some Data Derived from the Analysis of 289 Responses to the Items of the Picture-Test. All Proportions Corrected for Multiple-Choice Chance Selection.

Item No.	C.R.	pq	r_{it}	Item No.	C.R.	pq	r_{it}
205	4.8	0.2275	0.45	401	9.6	0.1824	0.63
206	5.4	0.2331	0.51	402	0.0(R)	0.1971	----
207	6.6	0.2499	0.42	403	7.8	0.2464	0.52
208	3.4	0.2331	0.24	404	3.6.	0.1411	0.18
209	4.0	0.2100	0.38	405	0.0(R)	------	----
210	4.3	0.1131	0.50	406	8.4	0.1600	0.70
211	5.2	0.2451	0.45				
212	3.2	0.2176	0.30	501	8.4	0.1771	0.72
213	3.6	0.2304	0.25	502	6.6	0.2451	0.51
214	0.0(X)			503	5.4	0.2451	0.40
215	0.6(X)			504	7.2	0.2139	0.55
216	0.0(X)			505	13.2	0.2499	0.76
217	4.8	0.1411	0.45	506	2.4	0.1875	0.25
218	0.0(X)						
305	6.0	0.0651	0.50	601	6.4	0.0736	0.60
306	3.6	0.1775	0.50	602	3.6	0.2331	0.28
307	10.2	0.2430	0.64	603	3.6	0.0099	0.40
308	0.6(X)			604	7.2	0.1600	0.63
309	1.6(X)			605	4.8	0.0475	0.52
310	1.6(X)			606	1.1(R)	0.2016	0.10
311	2.8	0.0979	0.32				
312	1.0(X)			701	1.0(R)	0.1875	0.11
313	4.6	0.2464	0.32	702	14.4	0.2491	0.81
314	2.2	0.2210	0.16	703	9.6	0.1875	0.70
315	2.4(X)			704	8.4	0.1875	0.65
316	0.0(X)			705	4.8	0.2419	0.34
317	4.9	0.2451	0.36	706	1.7(R)	0.1344	0.20
318	1.4(X)						
319	5.7	0.2176	0.44				
320	1.4(X)						
321	7.8	0.1771	0.60				
322	0.0(X)						

Key: C.R. is $P_U - P_L \,/\, \sigma_{P_U - P_L}$ where P_U is proportion of upper 25 per cent passing the item, and P_L is proportion of lower 25 per cent passing item; p is proportion of all pupils passing item; q is (1-p); r_{it} is item-test correlation; (X) shows that item was rejected; (R) that item was retained.

Pupil Score and Data Card.

E. G. High　Earl Godfrey　Boy　15 yr. - 6 mo.

Gen. Sci. - 3 yrs.　Sept. 9, 1930

Grade 9-

Otis I.Q. 110

I	II	III	IV
5	6	7	8
703	701	705	706
602	704	606	704
601	405	320 -	401
310	322 -	315 -	313
	318 -	218 -	214 -
	207	216 -	210
43			~~114~~
13			
28			

PT 28

AT 52

Test # 4

Table 16 . Data from the Three Tests.

X for Picture-test
Y for Achievement Test
Z for I.Q.'s

Total scores	X	6873	
Total squared scores	X^2	181419	
Mean		25.36	
σ_x^2		26.27	N = 271
$\sigma^2_{pop.}$		26.37	
σ_x		5.13	

Total scores	Y	13366
Total squared scores	Y^2	706554
Mean		49.32
σ_y^2		174.70
$\sigma^2_{pop.}$		175.34
σ_y		13.21

Total scores	Z	29534
Total squared scores	Z^2	3276659
Mean		108.98
σ_z^2		214.18
$\sigma^2_{pop.}$		214.98
σ_z		14.64

5958-19

BOSTON UNIVERSITY
1 1719 02570 7805